Structural Irrationality

Structural Irrationality[1]

T. M. Scanlon

Many normative claims are substantive claims about reasons— claims, for example, about the reasons that a person in certain circumstances has to do or to believe something. But not all normative claims are substantive claims about reasons. In particular, some claims about what it would be irrational for someone to do are normative claims but not claims about the reasons that person has. Here are some examples.

If a person believes that p, then it would be irrational for him to refuse to rely on p as a premise in further reasoning, and to reject arguments because they rely on it. To say this is not to say that the person has good reason to accept these arguments. Perhaps what he has most reason to do is to give up his belief that p. The claim is only that *as long as he believes that p*, it is irrational of him to refuse to accept such arguments. Similar claims hold in regard to practical reasoning: if a person intends to do A at t, and believes that in order to do this she must first do B, then it is irrational for her not to count this as a reason for doing B. This is not to say that she has any reason to do B. Perhaps what she has most reason to do is to abandon her intention to do A, or to change her mind about whether it is necessary to do B in order to do A later. But as long as she does not do either of these things, it is irrational for her to deny that she has any reason to do B. Normative claims of this kind involve claims about what a person must, if she is not

[1] I am grateful to participants in the discussion at the Common Minds conference for their helpful criticisms and suggestions. For comments on later versions, I am indebted to Luca Ferrero, Pamela Hieronymi, Nadeem Hussain, Niko Kolodny, Derek Parfit, Jay Wallace, and members of my Fall, 2003 Colloquium for first year graduate students.

irrational, treat as a reason, but they make no claims about whether this actually *is* a reason.[2]

I will call claims of this kind *structural* claims about rationality, to distinguish them from *substantive* claims about what is a reason for what. They are structural because they are claims about the relations between an agent's attitudes that must hold insofar as he or she is not irrational, and the kind of irrationality involved is a matter of conflict between these attitudes. In earlier work, I have suggested that we should restrict the term 'irrational' to instances of what I am here calling structural irrationality.[3] I am not relying on that restriction here. My present thesis is just that some claims about what a person must do insofar as he or she is not irrational are of this kind.

In this paper I will first examine in more detail the kind of normativity involved in requirements of structural rationality. I will then consider how these requirements are to be formulated, first in general and then with regard specifically to intentions and to beliefs. Finally, I will consider the implications that my conclusions have for the widely discussed idea that beliefs and desires have different "directions of fit."

The Normative Basis of Structural Rationality

What I am here calling structural requirements of rationality are examples of what Philip Pettit calls "programmed regularities that have the status of norms."[4] They are

---

[2] I believe that these are what John Broome has called normative requirements. See, for example, "Reasons," in *Reason and Value: Themes from the Moral Philosophy of Joseph Raz*, edited by R. Jay Wallace, Michael Smith, Samuel Scheffler, and Philip Pettit (Oxford: Oxford University Press, 2004), pp. 28-55.

[3] *What We Owe to Each Other* (Cambridge, MA: Harvard University Press, 1998), pp. 25-30.

[4] See p. 183 of his "Three Aspects of Rational Explanation," in *Rules, Reasons, and Norms* (Oxford: Oxford University Press, 2002), pp. 177-191.

regularities that any being will generally conform to insofar as it is a rational agent, and ones that can serve as the basis for explanations of such a being's behavior. We can explain the fact that a rational agent accepts a certain argument by referring to its beliefs, and we can explain what it does, or declines to do, by reference to what aims or intentions it has adopted, or judged there to be good reasons to adopt. These explanations are what Pettit calls "programming explanations" because the regularities they invoke hold in virtue of lower level causal regularities guaranteed by the physical states that realize the psychological states in question. To claim that some being is a rational agent is, in part, to claim that it is so constituted, physically, that these regularities will in general hold. It is also to claim that this being is one for which these regularities are norms. I want now to consider in what way such regularities are normative.

The behavior of a rational agent will exhibit (at least to a significant degree) the regularities described by requirements of rationality. But this is not because the agent sees this way of behaving as required by principles that she must be guided by. A rational agent who believes that p does not accept arguments relying on p as a premise *because* she sees this as required by some principle of rationality to which she must conform. Nor does she generally do it "in order not be irrational." Rather, she will be willing to rely on p as a premise simply because she believes that p. Similarly, a person who believes that doing A would advance some end of hers will not see this as counting in favor of A because some principle requires her to so count it, or because she must do this in order to avoid irrationality. Rather, insofar as she is rational she will see the fact that A would advance this end as a reason for doing A simply because she has the end in question. Ideas of rationality and irrationality belong to a higher-order form of reflective thought

that we need not engage in when, for example, we see that we have reason to take means to our ends.

Nonetheless, a person who violates these requirements can be described correctly by others as irrational, and she can so describe herself. Moreover, irrationality of this sort is a defect, a failure to meet standards that apply to us. John Broome puts this by saying that the requirements I have been discussing are ones that we *ought* to conform to. He says, for example, that "You ought (to intend to M if you intend to E and believe your M-ing is a necessary means to your E-ing)."[5] But how are such 'oughts' to be understood? What kind of normativity do they involve?

One might say, plausibly, that a charge of irrationality is a judgment of functional deficiency, of the same kind as a judgment that a carburetor, or a kidney, is deficient because it does not operate in the appropriate way. This is what Pettit suggests when he writes that "A regularity will count as a norm for a system just in case the satisfaction of that regularity is required for the system to succeed in the role for which it has been designed or selected."[6]

This sounds right, but there is a question as to whether this idea of a norm can account for the normativity involved in charges of irrationality of the kind we are considering. There are other modes of functioning, such as having a capacity and desire to reproduce, that we have been selected for and which therefore constitute norms for us in the sense that Pettit describes. It may be true in a functional sense that we "ought" to have the capacity to reproduce—that we are functionally defective if we lack this

---

[5] "Reasons," p. 29. Broome means the parentheses to indicate that the ought in question is of 'wide scope," and therefore non-detaching. That is to say, from S ought (to intend M if he intends E and believes that p), S intends E, and S believes that p, one cannot infer S ought to intend M.

capacity. But these norms, and 'oughts,' need have no normative force for an agent who recognizes them. I can recognize that, in this functional sense of the term, members of my species ought to reproduce, without believing that, in any sense that is even remotely action-guiding, I myself ought to reproduce.

I said earlier that a person who sees the fact that some action would advance her end as a reason to do it need not reach this conclusion by way of the idea of rationality or see this action as required by a norm that she is guided by. Nonetheless a person who sees that she has been irrational will see this as a defect in a sense that goes beyond the functional sense just described. She will see her attitudes as in need of revision—feel some "normative pressure" to revise them. So there is a question about how this kind of normativity is to be understood.

We can approach this question by considering what Pettit goes on to say in the article I have been discussing. He distinguishes there between what he calls normalizing explanations and interpretive explanations, which appeal to the fact that the subject of explanation saw things in a certain way. He observes that normalizing explanations need not be interpretive, and as an example cites decision-theoretic explanations. We might, he says, assign to a subject certain probability and utility functions and explain its behavior by seeing it as maximizing its expected utility (given these functions.) But in order to do this we need not suppose that the subject reasons in terms of those functions, or even that it is conscious at all. This would be a normalizing explanation (if, for example, we assume that the subject is designed or selected for utility maximization) but not an interpretive one. Utility maximization would be a norm for this subject only in a thin, purely functional sense.

---

[6] Op. Cit. p. 183.

It is also true that beliefs and preferences are attributed to the subject in this example only in a very thin sense. Indeed, one might question whether such a subject can be said to have beliefs or preferences at all. So let us enrich the example by supposing that the subject is a conscious agent and that it is designed or selected not merely to maximize its utility but to do this by reasoning in terms of probability and desirability. If this is so, then it would not only be a functional defect in the subject if it failed to maximize its utility but also if it failed to think about doing so in the proper way—for example, if it failed to see new evidence as a reason to revise its probability assignments in a certain way.

Would this give utility maximization normative force *for the agent* of the kind I have been discussing? For reasons I have already mentioned, the mere fact that the subject was selected to reason in terms of utility maximization would not make it more than a functional defect for it to fail to do so, or to do so in the right way. One way to ground a stronger claim would be to argue that insofar as the agent *sees* itself as reasoning in terms of probability and utility, it must see certain considerations as providing reasons. The claim would be that insofar as this is what it sees itself as doing, it cannot, insofar as it is not irrational, refuse to accept certain considerations as counting in favor of an action, or in favor of a change in belief.

This explanation has two components. First there is the "constitutive" claim that seeing things a certain way, or reasoning in a certain way, involves seeing certain things as reasons. This is a purely analytical claim. Normativity enters only from the point of view of the person who has these attitudes, and therefore sees the relevant considerations

as reasons. The "normative force" we have been trying to explain is just the force of those reasons, for the agent who sees them as reasons.

I want to argue that the normative content of requirements of rationality that I have been discussing has this same character. It lies in the fact that insofar as a subject has beliefs and intentions, it must see these as responsive to its assessment of the reasons for these states; and insofar as it has a certain belief, or intention, it must see this as providing the basis for further reasoning about what to believe and what to do. The relevant norms are thus elements of ("constitutive of") certain attitudes, and the relevant normativity is provided by what the agent sees as reasons.[7] I will now try to argue for this by spelling out in more detail how it might work in the cases of intention and belief.

Formulating Requirements of Structural Rationality

If there are rational requirements governing attitudes such as belief and intention, what is their content? So far, I have stated them as requirements dealing with what an agent who has certain attitudes must treat as a reason for certain other attitudes. For example, I have said that insofar as an agent believes that p, he or she must treat the fact that q follows from p as a reason for believing that q, and insofar as an agent intends to do A at t, he or she must treat the fact that doing B is required in order to do A at t as a reason for doing B. These formulations seem to me inadequate in several respects. As a start toward seeing why, we should ask what is meant here by "believing that p" or "intending to do A at t."

---

[7] In putting the matter this way I am following Niko Kolodny, "Why Be Rational?" *Mind* 114 (2005), pp. 509-563. I have learned a great deal from that paper and from discussions with Kolodny.

Believing that p can involve a number of different things: judging there to be sufficient evidence for the truth of p, being willing to (sincerely) affirm p, and being disposed to rely on p as a premise in further argument—that is to say, disposed to regard the fact that something follows from p as a reason to accept it as true. [8]

Having an intention to A at t can involve similarly diverse elements: judging that one has good reason to A at t, having consciously decided to do A, and being disposed to take one's doing A at t into account in further reasoning by, for example, treating the fact that doing B would facilitate doing A at t as a reason for doing B and treating the fact that doing B would interfere with doing A at t as a reason against doing B.[9]

All of the elements I have mentioned may be present in an ideal case of belief or intention, but they need not all be present every case. Someone might have no view about whether p is supported by the evidence (or even think that it is not), and might be unwilling to affirm that p when asked, yet might regularly rely on p as a premise in theoretical and practical reasoning. We might say of such a person that he believes that p even though he denies it and, if he judges there to be conclusive evidence against p, that he is irrational in so believing. On the other hand, if a person consistently refused to rely on p as a premise, and rejected arguments relying on it, then it would be plausible to say that he did not really believe that p even though he in fact judged p to be supported by conclusive evidence. (Readiness to affirm p is a swing case. If the person in the example just mentioned were unwilling to affirm that p, this would, I think count strongly in favor

---

[8] Peter Railton makes a similar point that belief and intention involve "bundles" of attitudes and dispositions which may sometimes come apart. See pp. 70-73 of his "On the Hypothetical and Non-Hypothetical in Reasoning about Belief and Action," in Cullity and Gaut, eds., *Ethics and Practical Reason* (Oxford: Oxford University Press, 1997).

[9] I will state my argument in terms of "intending to do A," but I believe that the same points could be made in terms of "having E as an end."

of saying that he did not believe that p. If he affirmed p but was unwilling to rely on it, this would be less clear.)

Similarly, even if a person judges himself to have good reason to do A at t, if he fails to give this factor any weight in further practical reasoning we might well say that he does not really intend to do A at t, even though he claims to have this intention and even though, given his assessment of the reasons for doing A at t, he may be irrational in not having it.

The fact that willingness to give a certain consideration weight in further theoretical or practical deliberation has this central a place in our criteria for attributing beliefs and intentions may lend support to the idea that readiness to reason in this way is "constitutive" of these attitudes, in the sense of being a necessary condition for having them at all. The problem that this raises for my present purposes is that it seems to turn the claim that, insofar as one has these attitudes, their contents must figure in one's subsequent reasoning in the relevant ways, into a tautology rather than a requirement of rationality that it is possible to violate.[10]

This problem could be avoided by identifying believing that p (or having an end, E) not with *actually* taking p and E into account in one's subsequent theoretical and practical reasoning but rather with being *disposed* to do so. On this reading, the requirements I have stated would not be tautologies, since a single failure to perform in a certain way does not show that an agent lacks the disposition to so perform. The problem with this interpretation of the requirements, however, is that they cease to be

---

[10] Christine Korsgaard emphasizes this as a problem for certain attempts to formulate the principle of instrumental rationality. See "The Normativity of Instrumental Reason," in Cullity and Gaut, eds, *Ethics and Practical Reason* (Oxford: Oxford University Press, 1997), pp. 215-254. I am indebted to her discussion.

requirements of rationality, since acting contrary to a disposition one has is not necessarily irrational.

So we need to look for a different way of understanding the content of these requirements. They specify that *if* an agent fulfills certain conditions, *then* he or she must, on pain of irrationality, have or not have a certain attitude. The question is how these antecedent conditions are to be interpreted. The hypothesis I will pursue in the rest of this paper is that these conditions consist in some judgment or commitment on the part of the agent. The first question to be addressed in carrying out this strategy is how this judgment should be interpreted.

One possibility, which I have sometimes invoked, is to interpret this antecedent as what I will call an attitude-directed judgment—a judgment about the adequacy of reasons for holding the attitude in question. So, for example, it might be that an agent who judges there to be conclusive reason to believe that p must, insofar as he or she is not irrational, believe that p. That is to say, he or she must be willing to affirm p, take the fact that q follows from p as a reason for believing q, and so on. Similarly, we could say that an agent who judges him or herself to have conclusive reason for intending to A at t must, insofar as he or she is not irrational, intend to do A at t, that is to say, must take the fact that doing B is necessary in order to A at t as a reason for doing B, take the fact that doing B would be incompatible with doing A at t as a reason against doing B, and so on.

One problem for this proposal is that the range of possible reasons for *having* a certain attitude may be too broad, because these could include "pragmatic" or "state-given" reasons for having a certain belief of intention.[11] For example, a person might

[11] I take the terminology of "state-given" and "object-given" reasons from Derek Parfit. See pp. 20-25 of his "Rationality and Reasons," in Dan Egonsson, Jonas Josefsson, Björn

have been promised a large reward for having that attitude or threatened with a terrible punishment for not having it. But these reasons might not provide grounds for giving weight to that belief or intention in further reasoning.

Leaving that problem aside for the moment, the formulations just given do seem to state genuine requirements of rationality. It does seem clearly irrational to have an attitude one that one explicitly judges oneself to have conclusive reason not to have. One might say that this irrationality just reflects the fact that belief and intention are what I have called judgment-sensitive attitudes, that is to say, attitudes that, insofar as we are rational, will be responsive to our assessments of the reasons for them.[12] It may be that non-human animals, and human infants, have beliefs and intentions that are not linked to assessments of reasons in these ways, because they not capable of making judgments about what they have reason to do or to think. But for us, who are capable of making such judgments, these connections hold. For us, belief and intention are attitudes that must, insofar as we are not irrational, be responsive to our assessments of relevant reasons. To say this is not to say that, for us, beliefs and intentions generally arise in response to conscious judgments about reasons. Clearly they do not. (Perceptual beliefs are obvious counter-examples, and there are many others.) To say that belief and intention are judgment-sensitive is only to say that *when we do make judgments about the relevant reasons* these attitudes will, insofar as we are rational, be responsive to them.

Even accepting that one's beliefs and intentions will, insofar as one is rational, be responsive to one's judgments, questions remain about which judgments in particular they are to be responsive to. So far, I have been describing them as attitude-directed

---

Petersson, and Toni Ronnow-Rasmessen, eds., *Exploring Practical Philosophy* (Aldershot: Ashgate, 2001).

judgments, which are explicitly about whether there are compelling reasons for the attitudes in question. This interpretation is what gave rise to the problem I mentioned about pragmatic reasons for having an attitude. But this problem reflects a larger difficulty, which is that judgments explicitly about the reasons for other judgments have a higher-order character that makes them somewhat artificial.

We do sometimes express what appear to be attitude-directed judgments. We might say, for example, that there is (or is not) good reason to believe that there are weapons of mass destruction in Iraq, and we may ask someone what his reason is for intending to take early retirement. But these may just be round-about ways of expressing content-directed judgments or questions, and a special context may be required in order for these attitude-directed formulations to sound natural. More commonly, when a belief or intention is arises from a conscious judgment, this judgment is content-directed. In deciding whether to believe that p we "direct our attention to the world" and ask whether p is true, and a judgment leading to an intention to do A at t is likely to be a judgment about the merits of doing A.

As I have remarked above, the rational requirements we are considering can be explained in a particularly direct way if the judgments that figure in the antecedents of these requirements are understood in attitude-directed form, since it seems clearly irrational to fail to have an attitude that one explicitly judges oneself to have conclusive reason for, or to continue to hold an attitude one explicitly judges oneself to have conclusive reason against. But if attitude-directed judgments are somewhat artificial then this explanation may not be one we should rely on. In any event, we should look for a

---

[12] See *What We Owe to Each Other*, pp. 20-24.

justification that would apply as well to the wider range of normal cases involving content-directed judgments.

There is also a potential problem of illicit generalization. The claim that it is irrational to fail to hold an attitude one judges to be supported by conclusive reasons (and irrational to continue to hold an attitude one judges there to be conclusive reasons against) is a perfectly general one. To recognize these as instances of irrationality one need not inquire into any features of the attitudes in question aside from their judgment-sensitivity. I have been inclined to rely on this generality, and to argue on this basis that belief and intention are more similar than commonly supposed. But insofar as charges of irrationality rest on the clash between particular attitudes and more specific content-directed judgments, the basis for these charges may be different for different attitudes. Belief and desire may both be judgment-sensitive attitudes, but the reasons why they are sensitive to particular content-directed judgments are, presumably, different. So we need to look more closely at those reasons.

Interpreting Rational Requirements: Intention

Consider first the case of intention. It will be helpful here to distinguish three attitudes: judging oneself to have conclusive (or sufficient) reason to do A at t, deciding to do A at t, and taking what is required to do A at t into account in one's subsequent assessments of one's reasons for doing, or not doing, other things. I do not mean to suggest, here, that every intention arises from a conscious decision. This is certainly far from being the case. But there is such a thing as deciding to do something (forming the intention of doing it), and this can come apart from the other attitudes I have listed. One

can judge oneself to have sufficient reason to do many different incompatible things, and one of the functions of deciding to do something is the necessary task of selecting among these. Nor is deciding to do something the same as judging oneself to have *conclusive* reason to do it. One can judge that one has conclusive reason to call the travel agent today to book a ticket, or to call the doctor today about the strange lump one has noticed in one's chest, yet not decide to do these things. Speaking for myself, I confess that I often fail to decide to do what I know that I have conclusive reason to do. Doing this is irrational but, sad to say, all too familiar.

So we have identified two points where irrationality is possible (that is to say, two points where there are normative links that can be violated.) The first is between judging oneself to have conclusive reason to do A at t and deciding to do A at t. The second is between deciding to do A at t and taking doing A at t into account in the proper way in one's subsequent reasoning. The task is to explain what kind of normative force the charge of irrationality has at each of these points.

If deciding to do A at t is something different from judging oneself to have reason, even conclusive reason, to do A at t, what more does it involve? Speaking metaphorically, one might say that it involves putting doing A at t on one's agenda, as something that is to structure one's further deliberation. Less metaphorically, it involves a commitment on one's part to think about what to do in a way that is compatible with one's doing A at t.[13] In particular, it is a commitment to take the fact that doing some

[13] It involves adopting a plan to do A in the sense described by Michael Bratman. See his *Intention, Plans, and Practical Reason* (Cambridge, MA: Harvard University Press, 1987). In my view, making such a commitment changes what one must, insofar as one is not irrational, see as a reason. But it does not give one a reason that one did not have before. On this see my "Reasons: A Puzzling Duality?" in *Reason and Value: Themes from the Moral Philosophy of Joseph Raz*, edited by R. Jay Wallace, Michael Smith,

action, B, would facilitate one's doing A at t as a reason for doing B, and to take the fact that doing B would be incompatible with one's doing A at t as a (normally conclusive) reason against doing B. It is irrational for someone who has decided to do A at t (and has not changed his or her mind about this) to refuse to treat the fact that B would facilitate this as a reason for doing B or to refuse to treat the fact that doing B would be incompatible with doing A at t as a reason against doing B. These things are irrational because they involve acting contrary to a commitment that one has made (and not revised.)

One can tell a functional story about why there should be such a thing as deciding to do something. We need to be able to do this because, in order to act effectively, we need to be able to structure our decision-making in certain ways. We often need to select among alternatives each of which is supported by sufficient, but not compelling reasons. And even when we have compelling reason for a particular alternative, in order to pursue that aim effectively we need to be able to give it a particular standing in our subsequent deliberation.

This functional story explains why there should be such an attitude as deciding to do something and why it should have the character that it does—why it should involve a certain kind of commitment. It is not yet, however, a full explanation of the irrationality

---

Samuel Scheffler, and Philip Pettit (Oxford: Oxford University Press, 2004), pp.231-246. Nadeem Hussain has pointed out in commenting on this paper that Bratman holds that what distinguishes intentions and plans from desires is that they are governed by consistency constraints and demands for means-ends coherence. (*Intentions, Plans and Practical Reason*, p. 31) It would therefore be circular, he says, to appeal to commitments (understood as what Bratman calls plans) in order to explain these constraints. Whatever Bratman's strategy may be, mine is not to explain commitment in terms of rational requirements. Rather, it is to explain these requirements in terms of a notion of commitment that I take to be understandable independently, in particular, in terms of the role such states play in our practical and theoretical thinking.

that is involved when a person, having decided to do A at t (and without revising that decision) refuses to treat the fact that B would facilitate his doing A at t as a reason for doing B. As I have said, the force of the charge of irrationality, for a person who finds himself in this situation, is not that of realizing that he has a certain functional deficiency. This force, it might be suggested, lies in the clash of attitudes that will be experienced by a person who has made a commitment of the kind described but finds himself with other attitudes that are incompatible with it. The importance of this clash is brought out by the fact that what is irrational is denying that the fact that doing B would facilitate my doing something I have decided to do is a reason to do B. Simply forgetting to do B, or absent mindedly overlooking the fact that one had decided to do A, would undermine the effective pursuit of my aim just as much as denying that I had reason to do what would promote it. But these failings would not be instances of irrationality, because they do not involve the appropriate clash of attitudes.

So we might say that the functional story explains why there should be such an attitude as deciding to do something and why it should involve a commitment of a certain kind, and that we then have a "constitutive" explanation of why someone who *has* a particular attitude of that kind and another attitude that it rules out should feel "normative pressure" to revise these attitudes.[14]

This is not the whole story, however. Insofar as the irrationality in question is just a matter of incompatible attitudes, we could avoid it by giving up either of them. But while it is true that refusing to treat the fact that doing B is necessary if one is to do A at t is irrational only so long as one has not abandoned one's decision to do A, it may seem

that the situations we are concerned with are not entirely symmetrical. If one has decided to do A at t, rationality may seem to speak more on the side of taking oneself to have a reason to do B than on the side of abandoning one's decision to do A. If there is this normative asymmetry, then something more is needed to explain it.

Insofar as there is an asymmetry here, however, it is qualified. If the steps necessary to do what one has decided to do turn out to be very costly, then what one has most reason to do may be to reverse that decision. Whether this is so depends on the reasons supporting the decision. What we are concerned with, however, is not what an agent has most reason to do but what he or she must do insofar as he or she is not irrational. For the purposes of answering *this* question, what matters is not the reasons that the agent actually has for doing A but, as Niko Kolodny has argued, the agent's assessment of these reasons.[15] If the agent's judgment is still that, even taking into account the cost of B as a necessary means, he has conclusive reason for doing A, then (whatever the merits of this judgment) it is irrational for him to resolve the conflict we have been discussing by abandoning his decision to do A. (This is the asymmetry we have been considering.) But if this is not his judgment, then abandoning A is not irrational (although it may still be inadvisable), and if he holds the opposite assessment, then it would be irrational for him *not* to abandon the decision to do A.

Earlier, I distinguished two points at which irrationality may occur (that is to say, two points where there are normative links that can be violated.) The first was between judging oneself to have conclusive reason to do A at t and deciding to do A at t. The

---

[14] The constitutive claim being just that deciding to do A at t involves committing oneself to giving one's doing A at t a certain place in one's subsequent reasoning about what to do.

[15] See Kolodny, "Why Be Rational?"

second was between deciding to do A at t and taking doing A at t into account in the proper way in one's subsequent reasoning. I have so far been addressing the second of these links. In explaining the irrationality of denying that one has reason to take steps necessary for doing what one has decided to do, I appealed first to a constitutive claim about deciding: because deciding involves seeing oneself as making a certain kind of commitment, an agent who has decided must see a conflict between this decision and his subsequent denial. Then, in order to explain why it is sometimes irrational for an agent to resolve this conflict by abandoning the decision, I had to appeal to the idea that there is a normative link between an agent's decision to do A and his assessment of the reasons for doing A.[16] But this is just the first of the two links that I earlier distinguished. So I must turn to the question of what this link involves and how it is to be explained.

First, if one judges oneself to have conclusive reason not to do A at t, it is irrational to decide to do A at t, or if one has decided this, not to reverse that decision. This is irrational because deciding to do A at t involves committing oneself to giving the needs of doing A at t a certain positive weight in one's practical thinking—to take one's doing A as something that can provide reason to do or not to do other things. But it is irrational to do this if one judges there to be conclusive reason against doing A at t.

Second, if one even judges oneself not to have sufficient reason to do A at t, it is irrational to give positive weight in one's practical thinking to what is required in order for one to do A at t, since one would then be giving one's doing A at t a weight that one judged it not to have. So if one judges oneself not to have sufficient reason to do A at t,

---

[16] In appealing to these two elements I am retracing, in slightly different way, the steps of Korsgaard's "The Normativity of Instrumental Reason." Her account of the instrumental principle appeals to a constitutive claim about the attitude of "having an end." But she

then it is irrational to decide to do A at t and, if one has so decided, irrational not to reverse this decision.

Finally, if one judges oneself to have conclusive reason to do A at t, is it irrational not to decide to do A at t? Well, perhaps not, if one's not deciding is simply a matter of forgetting, or falling asleep, and not if one believes that A is something one cannot do.[17] Is it, then, at least irrational to *decline* to decide to do A at t, or to decide not to do it even though one believes one could? To do either of these things would be to consciously decline to take account of a consideration (one's doing A at t) that one in fact judges to be significant. There is, however, a temporal factor here that may be important. If t is some time far in the future, one might judge that one has conclusive reason to do A at t but no reason to take this into account in one's present thinking about what to do. So declining to decide, now, to do A at t would not be, as I just said, to consciously ignore what one judged to be a significant consideration. If one can costlessly defer the decision, then doing so would not be irrational, and it might even be favored by a principle of economy of thought (not encumbering one's deliberation with unnecessary commitments.)

One might summarize this line of thought by saying that even if one judges oneself to have conclusive reason to do A at t, it is irrational to decline to decide (now) to do A at t only if such a decision is needed in order to facilitate one's doing A at t. This might be taken to suggest that declining (now) to decide to do what one judges oneself to have conclusive reason to do later is irrational only when deciding now to do it would facilitate one's doing it then, and thus that this is irrational only when it is a violation of

---

also argues that the normativity of this principle cannot be explained without appeal the normative standing of the agent's end itself.

[17] For this last qualification I am indebted to Alison McIntyre's discussion in her unpublished paper, "What's Wrong with Weakness of Will?"

instrumental rationality. This might seem to threaten a regress, insofar as what we are trying to explain is in part the normative force of requirements of instrumental rationality. In addition, it seems to me artificial to describe one's decision to do something as a means to doing it. I therefore would prefer an explanation that appeals only to the bare idea of irrationality as failing to give a consideration the weight in one's thinking that one in fact judges it to have. If matters that are the subject of current deliberation will affect one's doing A at t, then declining to decide (now) to do A at t will involve irrationality of this kind. But if t is so far in the future that nothing in one's current deliberations will bear on it, then in failing to decide, now, to do A at t one will not be failing to give one's doing A then the weight that one judges it to have. So no irrationality will occur.

For an agent, the force of the three normative links between an assessment of the reasons for doing A at t and a decision to do A at t lies in the incompatibility that the agent who violates these links must feel between her various normative attitudes. As I noted previously, however, mere incompatibility of attitudes alone is symmetrical: it can be avoided by giving up either attitude. But it would be irrational for an agent to avoid the incompatibility between judging herself to have compelling reason to do A at t and her not deciding to do this by abandoning the former judgment unless she saw some reason to revise this assessment. And it is difficult to imagine a case in which she could take her failure to decide to do A at t as a consideration bearing on the merits of doing it.

Finally, what I have said here does not assume or imply that a rational agent holds the view that she *ought* to decide to do what she judges herself to have conclusive reason to do, or that he *ought* to give doing the things he has decided to do a certain place in his

subsequent reasoning. My claims have been only about what an agent, insofar as he or she is not irrational, will see as reasons.

Interpreting Rational Requirements: Belief

Let me turn now to the case of belief. Proceeding in parallel with the case of intention, we might distinguish the following three things that can be true of a person:

(1) S judges there to be conclusive evidence for the truth of p.

(2) S accepts p as true.

(3) S accepts p as a premise in further theoretical and practical reasoning (that is to say, S takes the fact that q follows from p as a reason for accepting q as true.)

In (2), "accepting p as true" is meant to be the theoretical analog of deciding to do A (or adopting A as an aim.) As I argued above, to decide to do A is to give doing A a certain status in one's practical reasoning, the status of something that provides reasons for doing what will facilitate this. Similarly, the idea would be here that to accept something as true involves giving it the status of something that is to be relied on in further theoretical reasoning by providing reasons for accepting what it entails, and to be relied on as a premise in practical reasoning.

As David Velleman has pointed, out there are various ways in which one can treat something as true. When we imagine that p, or accept p as a hypothesis or an assumption in order to see what follows from it, we are, in a sense, "treating p as true," and at least in the latter case we are showing a disposition to rely on p in subsequent reasoning (albeit reasoning of a hypothetical sort.) But, as Velleman says, believing that p involves more

than this: it "entails regarding p as 'really' true."[18] I intend "accepting p as true" to be understood in this stronger sense.

One difference between the cases of belief and intention should be noted at the outset. One important function of deciding to do something is to select among various alternative courses of action each of which one has sufficient reason to do, and to identify the one that is to form the basis of one's subsequent decision-making. In the case of belief, when one has incomplete information one may have to choose which of a set of plausible hypotheses to rely on in deciding what to do. But the result of this kind of selection is not *belief*.

Another difference between the cases of belief and intention, is that in the case of belief it may be questioned whether (2) could be true of a person without (3) also being true. Could someone accept that p is something that is to be relied on in these ways yet refuse on some occasions to so rely on it? It seems to me that this is possible, but this may be less clear than in the practical case. In any event, I do not need to argue for this, since it is at least clear that if this were to occur it would be an instance of structural irrationality of the kind I have been discussing. Being (really) true is sufficient to give something the status it needs to be a premise in further reasoning. So accepting p as true while refusing to accept arguments that employ it as a premise would involve failing to recognize something as having the status in one's reasoning that one has acknowledged it to have.

I will therefore set aside the question of the normative link between (2) and (3) and concentrate on the relation between (1) and (3): between a person's assessment of the

[18] David Velleman, "The Guise of the Good," *Nous* 26 (1992) pp. 3-26. The quoted passage is from p. 15. Velleman says that this entails "regarding p not only as true but

evidence for p and one's willingness to accept arguments that employ p as a premise—that is, to take the fact that q follows from p as reason to accept q as true.

Velleman suggests that we should take a link between these as definitive of belief, but the link he has in mind is dispositional. He writes that "when someone believes a proposition … his acceptance of it is regulated in ways designed to promote acceptance of the truth; he comes to accept the proposition, for example, when evidence indicates it to be true, and he's disposed to continue accepting it until evidence indicates otherwise. Part of what makes someone's attitude toward a proposition an instance of belief rather than assumption or fantasy, then, is that it is regulated in accordance with epistemic principles rather than polemics, heuristics, or hedonics. An attitude's identity as a belief depends on its being regulated in a way designed to make it track the truth."[19]

But this does not seem right. It does not seem necessary, in order for an attitude to count as belief, that the subject is *actually* disposed to regulate it in a way designed to make it track the truth. We may have some beliefs that we are careful to screen off from any critical assessment of the evidence for or against their truth. In so doing, we are irrational (especially if we do so because we think that the evidence is probably overwhelmingly against these beliefs.) But we still, I would say, believe these things. Velleman's proposal seems to make this kind of irrationality impossible.

Perhaps one should say that an attitude counts as belief only if the subject recognizes that he *should* regulate it in this way. Lloyd Humberstone makes a similar suggestion. He says that "unless one takes there to be a criterion of success in the case of an attitude toward the proposition that p, and, further, takes that criterion to be truth, then

---

also correct to regard in this way." I will return to this part of his view.

[19] "The Guise of the Good," p. 14.

whatever else it may be the attitude in question is not that of belief. So unless the attitude-holder has what we might call a controlling background intention that his or her attitudinizing is successful only if its propositional content is true, then the attitude taken is not that of belief."[20]

So, following Humberstone, we could say that a person "accepts p as true" in the sense intended in (2) and (3) only if he or she has a controlling intention of this kind. It follows that a person must see this attitude as one that ought to be responsive to what he or she takes to be evidence for the truth of p. This provides a link between (1) and (2) that grounds a requirement of structural rationality. If a person regards "acceptance as true" as an attitude that is successful only if p is "really" true, then she will regard this attitude as one that should be formed if there is conclusive evidence for the truth of p and abandoned if there is conclusive evidence that p is false[21]

There is a question here about what might be called the "initial direction" of the controlling intention that Humberstone describes. Velleman proposed that in order for a state to be a belief, the agent had to be disposed to modify it in the light of what he or she took to be evidence of its truth. So he proposed a dispositional link reaching, so to speak, backwards from (2) and (3) to (1). One might take Humberstone to be proposing a link in the same direction, but one consisting of an intention rather than a disposition—an intention to regulate one's belief in the light of one's assessment of the evidence for its truth. This would provide a link of the right kind between (2) and (1): one that can be violated, but only on pain of irrationality. An agent who accepted p as true despite judging that there was conclusive evidence against its being true, or who refused to

---

[20] I. L. Humberstone, "Direction of Fit" *Mind* 101 (1992) 58-83, p. 73.
[21] As before, I leave aside here the possibility of "state given" reasons for having a belief.

accept p as true although he also judged there to be conclusive evidence in its favor, would be irrational because he would be failing to regulate his acceptance of p in the way required by his own "controlling intention."

This is quite plausible. It does, however, depend on a higher-order, attitude-directed intention. One may wonder whether an agent must have such an intention with regard to anything that can be called a belief, even beliefs that one holds in the face of what one sees as contrary evidence. A person who has such a belief seems in some way committed to modifying his belief in the light of evidence, but this commitment may not be best expressed in terms of an attitude-directed intention to do just this.

An alternative account of this link would take Humberstone's controlling intention as facing in the opposite direction: from (2) toward (3). The idea would be that a person accepts p as true in the way involved in belief only if he or she intends to rely on p in further reasoning about what the world is like, and about what to do. A belief held with this intention is "successful" only if it is appropriate to rely on it in these ways—that is, only if it is true. Although it does not mention grounds or evidence, this intention brings with it a rational requirement linking (2) and (1). It is irrational to take p as having the status just described while simultaneously judging there to be conclusive evidence against the truth of p. (Similarly, it would be irrational to refuse to accept p as a premise in further argument if one judges there to be conclusive evidence for the truth of p.) This irrationality consists simply in the conscious holding of attitudes that are directly incompatible. The normative pressure that an agent who is irrational in this way feels to modify his acceptance (or non-acceptance) of p comes not from his acceptance of some

higher-order intention or "ought" judgment, but simply from the reason-giving force that he attributes to the evidence against (or for) the truth of p.

Concluding Thoughts about "Direction of Fit"

The points made in preceding sections about intentions and beliefs have some bearing on the idea that beliefs and desires are distinguished by having different "directions of fit." In this concluding section I want to examine these implications.

I take claims about reasons, such as "I have conclusive reason to A at t" to be ordinary declarative statements that can be true or false and can be the objects of belief.[22] I have been arguing that a person who believes that she has conclusive reason to do A at t will, insofar as she is not irrational, intend to do A at t and (absent change of mind, irrationality, or incapacity) will do so. It may seem, then, that at least on one understanding of the idea of a "direction of fit," the belief that one has conclusive reason to do A at t has, on my view, two directions of fit. As a belief, it is something that is defective if it does not "fit the world" (that is, if one does not in fact have conclusive reason to do A at t.) On the other hand, it is something that "the world must fit" (that is, insofar as one accepts it, and is not irrational, one will undertake to bring it about that one does A at t.) My view may therefore seem to be in tension with Michael Smith's arguments that there cannot be a state (what he calls a "besire") that has both "belief-like" and "desire-like" like directions of fit.[23]

The idea of different "directions of fit" can be given either a normative or a more descriptive reading. As Smith first states it in *The Moral Problem*, the idea is presented in

---

[22] I argue for this in "Metaphysics and Morals," *Proceedings and Addresses of the American Philosophical Association* 77 (2003).

what sounds like normative terms: a belief is a state that *must* fit the world, whereas a desire is a state that the world *must* fit. The normative ring of these 'must's is explicit in the passage that Smith quotes from Mark Platts.[24] Platts writes that "falsity is a decisive failing in a belief, and false beliefs should be discarded; beliefs should be changed to fit with the world, not *vice versa*." But, by contrast, "the fact that the indicative content of a desire is not realized in the world is not yet any reason to discard the desire; the world, crudely put, should be changed to fit with our desires, not *vice versa*."

When the distinction is put in this way, it is clear that a state cannot have both directions of fit with respect to the same content. A state cannot be both a belief that p and a desire that p, since a belief that p is defective and should be withdrawn of p is not the case, but this is no fault in a desire that p.[25] But it does not follow that a single state could not have two different directions of fit with respect to different contents: that, say, a desire that p could not be a state that the world must fit (with respect to p) but at the same time a state that is defective, and ought to be revised, if it fails to fit the world in some other respect q (such as that there is reason to bring about p.) Indeed, the position that Smith himself goes on to defend in *The Moral Problem* seems to be that desires have essentially this character.

This does not mean, however, that desires have both "directions of fit" as Smith understands this notion, because he goes on to define the idea of a direction of fit in dispositional rather than normative terms. He first defines desires and beliefs in terms of their functional role. "Under this conception," he writes, "we should think of desiring to

---

[23] See *The Moral Problem* (Oxford: Blackwell Publishers, 1994) pp. 118-125.
[24] *The Moral Problem*, p. 112. The quoted passage from Platts is from his *Ways of Meaning* (London: Routledge and Kegan Paul, 1979) pp. 256-257.
[25] See *The Moral Problem*, p. 118.

φ as having a certain set of dispositions, the disposition to ψ in conditions C, the disposition to χ in conditions C′ and so on, where in order for conditions C and C′ to obtain, the subject must have, *inter alia*, certain other desires and also certain means-ends beliefs, beliefs concerning φ-ing by ψ-ing, φ-ing by χ-ing and so on."[26] The idea, then is that to desire to φ is to be disposed to do those things that one takes to be ways of φ-ing.

Smith does not give a parallel account of the functional role of belief, but I assume it might be something like the following: We should think of believing that p as having a certain set of dispositions, such as a disposition to affirm p under certain conditions C, to affirm that q under conditions C′ (which include the subject's believing that p entails q), to affirm r under conditions C″ (which include the subjects believing that p entails r), and so on.

I have no objection to this. I have suggested above that believing and desiring involve dispositions of the kind Smith describes, from which it follows that on my account intending to do A involves desiring to do A in Smith's broad functional role sense of desiring. On both his account and mine, believing that p involves more than the dispositions just listed. And on my account (and I think Smith might agree) intending does as well.

The additional element of belief that is important on Smith's account is a higher-order disposition to modify the disposition I have listed. He writes that, "a belief that p tends to go out of existence in the presence of a perception with the content that not-p." But he holds that a desire that p does not tend to change in this way, and this difference in counterfactual dependence is the difference in "direction of fit."[27] Having stated the

[26] *The Moral Problem*, p. 113.
[27] *The Moral Problem*, p. 115.

notion of "directions of fit" in this dispositional rather than normative form, Smith then uses it as the basis for an extended argument against the possibility of states with both directions of fit—an argument that does not apply only against a state's having both directions of fit with respect to the same content.

This is what might be called a modal separability argument, which Smith states as follows: "[I]t is always at least possible for agents who are in some particular belief-like state not to be in some particular desire-like state; … the two can always be pulled apart, at least modally."[28] Smith's main example of separability is a case of someone who believes that a certain action would be morally right but has no desire (no relevant dispositions) to do it. But, in line with the discussion earlier in this paper, we can put the point more generally, in terms of someone who judges that she has conclusive reason to do A at t, but who fails to be disposed to do A at t or to do the things that she believes are necessary if she is to do this. I have argued at some length that this is possible, so I am here in firm agreement with Smith. There are, however, a few points that I would add.

The first is that the phenomena just described, when they are not the result simply of such things as forgetfulness, or falling asleep, or a blow on the head, are cases of structural irrationality. Here Smith would, I think, agree. Much of Chapter 5 of *The Moral Problem* is devoted to arguing for the claim

> C2 If a person believes that she has a normative reason to φ then she rationally should desire to φ.[29]

Although Smith and I agree on this claim, the interpretation and defense that I have provided for it differs from the one he offers in *The Moral Problem*.[30] On his

---

[28] *The Moral Problem*, p. 119.

interpretation, to believe that one has normative reason to φ is to believe that one would desire to φ if one were fully rational. Given this interpretation of a belief about one's normative reasons, C2 is then explained via the idea that a rational person's desires will conform to what that person judges to be required by rationality. As Smith says later in a slightly different context, but one that I take to be relevant, "an evaluative belief is simply a belief about what would be desired if we were fully rational, and the new desire is acquired precisely because it is believed to be required for us to be rational."[31] So on his interpretation the belief that one has reason to φ not only is what I have called above "attitude-directed" but also works explicitly via the agent's notion of what is required in order to be rational.

On my interpretation, by contrast, a judgement about one's reasons for doing A is just that—an assessment of the strength of certain reasons. It may entail a belief about what one would believe or do if one were fully rational, but it is a judgment with distinct content. As I made clear above, I allow for the possibility that judgments about what one has conclusive reason to do or believe may be attitude-directed. I began this paper with this interpretation in mind, and I think it is appropriate in some cases. But, as I have said, I believe that these are only some of the cases, and that in other cases structural rationality requires one to form certain beliefs and intentions in response to content-directed judgments about the reasons one has. Finally, I have maintained that agents, insofar as they are rational, will form certain beliefs and intentions (that is to say, acquire certain complex dispositions) in response to their judgments about the reasons they have,

---

[29] *The Moral Problem*, p. 148. I assume that Smith is here identifying "desiring to φ" with having the complex set of dispositions he mentions.
[30] I do not know whether Smith's current view has the features I am here discussing.
[31] *The Moral Problem*, p. 160.

and that they will see these attitudes as supported and even required by the contents of those judgments. But it seems to me that in most cases (I am tempted to say, in normal cases) this process will not involve explicit reference to their ideas of rationality: they will not form these attitudes because they see them as required by rationality or required in order to avoid irrationality.

I will mention two further points, which may be mainly terminological. The first is that Smith identifies having the dispositions that characterize desire with "having a goal." Thus he says, ""since all there is to being a desire is being a state with the appropriate direction of fit, it follows that having a goal just is desiring."[32] Having the dispositions that on Smith's view amount to a desire that p may amount to having a goal in a very thin sense, in which a person has the goal p just in case he or she is disposed to promote p. But if someone has what I would call an aim then it is not only true that she is disposed to pursue this aim but also that she is irrational if she fails to take the fact that something will advance that aim as a reason for doing it. Merely having a disposition does not provide a basis for this charge of irrationality, since it is not irrational to fail to act on a disposition one has.

The second point is that it seems to me misleading to use the terms 'belief-like' and 'desire-like' to refer to the kind of states that Smith's modal separability argument claims can always be pulled apart. As I have argued above, our common understanding of belief, like that of practical attitudes such as intention, involves diverse elements. Intending to do A at t generally involves both judging oneself to have sufficient reason to do A at t and being disposed to take one's dong A at t into account in one's thinking about what to do. Having E as an aim generally involves both judging E to be worth

pursuing and being disposed to take this aim into account in thinking about what to do by, for example, taking the fact that doing B would advance E as a reason for doing B. Similarly, believing that p generally involves not only judging there to be good evidence for the truth of p but also having various dispositions to rely on p in further reasoning. In each case, these elements can be "pulled apart:" one can judge oneself to have conclusive reason to believe that p, yet fail to have the relevant dispositions to rely on p in further reasoning, just as one can judge oneself to have conclusive reason to do A at t yet fail to take the fact that doing B is necessary to one's doing A at t as reason to do B.

So the phenomena to which Smith's modal separability argument calls attention occur in the case of belief as well as in that of attitudes like intention and having an aim. What can be "pulled apart" are not "belief-like" states and "desire-like" states but, rather, the distinct components of many states, including both beliefs and intentions.

---

[32] *The Moral Problem*, p. 116.